uracle

AURDA

엔터프라이즈 AI 운영의 완성

2026

table of content

Uracle is : 안정성 및 성장성

유라클은 모바일 플랫폼을 비롯한 다양한 업종과 분야에서 시스템 구축 사업을 수행하고 역량을 축적해 온 25년차 전문기업입니다.

uracle

AI 플랫폼·LLM·자체 개발 기술력 보유

회사명	㈜유라클	대표자	조준희, 권태일
주소	서울시 서초구 방배로 231 (방배동)		
사업분야	AI 솔루션 개발, 시스템 통합 구축·유지, 모바일 시스템 구축·개발		
설립년도 / 사업기간	2001.01 ~ 2026.07 (25년 6개월)		
보유인력현황	총 254명 (개발:218/지원인력:36)		

'25 신용평가등급
A

최근 3년 평균 매출
350억+

2001 ~ 2009

- '08 벤처기업대상 중소기업청장표창 수상
- '07 ㈜유라클 상호변경
- '07 정보통신부 핵심동력개발사업 차세대기술선정
- '01 '모바일로' 금융감독원 금융 신상품 대상
- '01 ㈜아이엠넷피아 설립

2010 ~ 2019

- '19 SW산업발전유공자 대통령 표창 수상
- '19 대한민국 소프트웨어 기업경쟁력 대상
- '16 SW기업경쟁력 대상 IT솔루션 분야 최우수상
- '15 모바일 전자정부 행정자치부 장관 표창
- '11 스마트앱어워드2011 금융서비스 최우수상

2020 ~ 2025

- '25 생성형 **AI 플랫폼 출시** (Athena AI 플랫폼)
- '25 **AI 코드어시스턴트 제품 출시** (Athena Code)
- '24 **고려대 HI AI 공동연구소 설립 및 LLM 개발**
- '24 ㈜유라클 KOSDAQ 상장 KOSDAQ
- '21 **한국소프트웨어산업협회 회장 취임** (조준희 회장)

사업 분야

AI 플랫폼 · 모바일 등 SW 개발,
시스템 구축 · 운영 · 유지보수

주요 제품

Athena — AI Product Suite

Morpheus — Mobile Product Suite

주요 고객사

- 대기업 그룹사

SAMSUNG 삼성전자 · 삼성생명 · SAMSUNG 삼성증권

HYUNDAI MOTOR GROUP · HYUNDAI STEEL · HYUNDAI GLOVIS · HYUNDAI AutoEver · HYUNDAI · HYUNDAI MOBIS · KIA

- 도메인 (일반 IT기업, 금융권 등)

KB · 신한카드 · 우리은행 · MIRAE ASSET · Hyundai Capital · Hyundai Card

- Other Industries

posco · SK hynix · SK innovation · SK telecom · CJ · kt · Hanwha

Uracle is : 고객사 현황

1,000여개의 산업 별 전문 지식과 경험으로 유라클이 AI 도입 여정의 새로운 파트너가 되겠습니다.

Uracle is : AI 사업 수행 실적

2025년 1월부터 AI사업 시작 후 상반기 20건의 PoC를 거쳐, 상용 레퍼런스를 빠르게 확장하고 있습니다.

사업명	발주처	투입 S/W	주요내용
인적 자율성 결합형 피지컬 AI 기반 트랙터 변속기 자율제조·검사 통합 플랫폼 개발	산업통산부/LS엠트랙	AI Platform	HITL(Human-in-the-Loop)/ 피지컬AI 트랙터 변속기 자율제조 통합 플랫폼 개발
고감도 누수센서를 활용한 생성형 누수감시 AI 플랫폼 실증 및 상용화	환경부/수자원공사,중부발전	AI Platform	생성형 누수감시 AI 플랫폼 실증 및 상용화
KOSA AI 서비스(KOSA-AX)	KOSA (한국인공지능·소프트웨어산업협회)	AI Platform	홈페이지 및 임직원 채널과 연동하여 경력·산업정보 조회·요약 및 시스템 연계형 업무지원 챗봇 구축
고효율 AI반도체 추론환경 및 플랜트 산업데이터 기반 LLM & 3D모델링 통합플랫폼 서비스 해외실증	과학기술정보통신부/사우디 아람코	AI Platform	플랜트 산업데이터 기반 3D 모델링 통합 플랫폼 해외 실증
AI 개발플랫폼 구축	KODATA(한국평가데이터)	AI Platform, AI Code Assistatnt	AI 개발플랫폼(LLMOps) 및 지식관리 시스템 구축
현대건설 AI 분양상담사 구축	현대건설	AI Platform	AI 분양상담 챗봇 대고객서비스 구축, 하반기 확산 사업 예정
AI 서비스 플랫폼 구축	KCB(코리아크레딧뷰로)	AI Platform	No code 기반의 전사 AI 서비스 구축용 플랫폼 구축
협회 통합정보시스템 구축	대한의사협회	AI Platform	AI 플랫폼 기반 민원인·직원·광고심의 신청자용 AI 챗봇 구축
AI기반 FDS 검사시스템 구축	우리은행	AI Platform	과거 사례 시나리오 기반 FDS 검사시스템 구축
AI 코드 어시스턴트 도입	KT 스카이라이프	AI Code Assistant	코드어시스턴트를 통한 IE/JAVA기반 구형 웹서비스 전환
AIOps 플랫폼 및 고객 지원 챗봇 개발	인텔리안테크	AI Platform	생성형 AI 플랫폼 및 고객 지원용 RAG 챗봇 구축
AI 서비스 내재화 플랫폼 구축	현대건설	AI Platform	전사 생성형 AI 플랫폼 구축
新 모바일 워크플레이스 구축	GS건설	AI Platform	전 임직원용 AI 비서(기사 요약, 앱 검색) 구축
생성형 AI 시범과제 구축 용역	메리츠증권	AI Platform	실시간 문서 번역/요약 Agent 구축
AI기반 차세대 안전신고 및 위험징후 분석 서비스 기술 개발	행정안전부	LG AI 연구원 EXAONE LLM	신고 Agent 개발 및 차세대 안전신문고 시스템 연동
지식 검색용 RAG 챗봇 PoC	한국투자저축은행	AI Platform	내부직원용 지식 검색 RAG 구축
홈페이지 고객용 RAG 챗봇 PoC	한국인공지능·SW산업협회	AI Platform	고객용 지식 검색 RAG 구축
경력관리 지식검색용 챗봇 PoC	한국건설기술인협회	AI Platform	경력 관리용 RAG 검색 테스트
ERP 매뉴얼 지식검색 PoC	두산퓨얼셀	AI Platform	ERP 매뉴얼 검색용 RAG 구축
품질검사 AI Assistant PoC	두산에너빌리티	AI Platform	검사원 선정 업무 자동화, 검사 체크리스트 생성

AI 인프라 운영의 과제

지금 필요한 것은 복잡한 인프라를 제어할 수 있는 새로운 운영 체계입니다.

기존 인프라 관리 방식의 한계

생성형 AI 확산으로 모델, 데이터, GPU·NPU 등 관리 포인트가 급격히 늘었습니다.
그러나 기존 인프라 운영은 여전히 서비스 중심으로 설계되어 있어, 복합 워크로드의 자원 제어와 최적화가 어렵습니다.

기업의 53%가 고밀도 AI 인프라 운영 전문 인력 부족
Flexential (2024)

인프라 자원의 다변화 및 효율적인 관리 필요성 대두

GPU는 핵심 자원이지만, 고비용·전력 문제로 인해 **NPU 등 추론 전용 HW 도입**이 대두되고 있습니다. 그러나 **이종 하드웨어 환경이 혼재**되며 통합 자원 관리와 LLM 서빙 최적화가 어려워, 고가의 자원이 유휴 상태로 남거나 효율이 저하되는 문제가 발생합니다.

AI 인프라 자원 평균 활용률 50~70% 수준으로 나머지는 유휴 상태
The State of AI Infrastructure at Scale (2024)

엔터프라이즈 환경의 복잡성

기존 플랫폼도 인증·보안을 제공하지만, 서비스별로 분리돼 통합 거버넌스가 어렵습니다.
AI 확산으로 데이터 접근·모델 호출·API 연동 등 관리 포인트가 늘며 엔터프라이즈급 통합 인증·감사·보안 체계가 필수가 되었습니다.

AI 워크로드를 PaaS 위에 그대로 올렸을 때 운영 복잡도가 평균 2배 이상 증가
Nutanix (2023)

AI 인프라 운영의 핵심 기술

새로운 체계의 핵심을 위해 자원·배포·가시화·보안 영역의 유기적 통합이 필요합니다.

구분	핵심 기능	요소 기술	기대 효과
통합 자원 관리	GPU, NPU, CPU 등 이종 하드웨어의 클라우드 중립적 추상화를 통한 연산 자원의 효율적 활용 및 자동 스케일링	GPU/NPU Operator 자원 스케줄링 분산 스토리지 서비스 이종 HW 풀링 기술	**자원 활용률 극대화** **운영비 절감**
배포 및 자동화	코드 변경부터 배포까지 전 과정을 자동화하는 통합 파이프라인	GitOps 자동배포 지능형 트래픽 라우팅 이미지 레지스트리 관리 올인원 서빙 게이트웨이	**배포 속도 향상** **무중단 운영 실현**
운영 가시화 및 복구	메트릭·로그·트레이스 및 HW 물리 지표(전력), LLM 추론 지표(TTFT/TPS)까지 운영 지표의 통합 수집을 통한 가시성 제공	메트릭 /로그 수집 트레이스 분석 HW 물리 모니터링 LLM 성능 지표 추적	**장애 사전 대응** **운영 효율성 향상**
지능형 자율 운영	MCP(Model Context Protocol) 기반 AI 에이전트가 자연어 명령을 해석하여 배포, 스케일링, 재시작 등 인프라 업무를 자율 수행	MCP 표준 프로토콜 인프라 제어 에이전트 자연어 명령 처리(NLU) 자율 오케스트레이션	**운영 복잡성 최소화** **업무 자동화 구현**
인증 및 거버넌스	서비스 전반의 인증·인가·비밀 관리·감사 로깅을 통한 통합 보안 체계	통합 인증·인가 정책 기반 접근 제어 감사 로그 및 추적 보안 정책 관리	**보안 강화** **규제 대응**

AURDA, AI 인프라 운영의 표준을 정의하다

운영 효율화

자동화된 자원 관리와 배포 체계로 운영 부담을 최소화하고, GPU·NPU·스토리지 활용률을 극대화합니다.

운영 가시성 및 자원 최적화

메트릭·로그·트레이스를 통합 관리하여 장애를 예측하고, 실시간 인프라 현황을 한눈에 파악할 수 있습니다.

엔터프라이즈 보안 관리

표준화된 인증·인가·감사 로깅으로 조직 전반의 서비스 안정성과 보안 준수성을 보장합니다.

확장 가능한 인프라

멀티 클라우드·온프레미스 환경을 아우르는 유연한 확장성으로, 기업의 성장과 변화에 민첩하게 대응합니다.

AURDA, 핵심 특장점

이기종 하드웨어부터 AI 에이전트 연동까지, 단일 플랫폼에서 구현하는 고효율·무중단 AI 운영체계 입니다.

이종 하드웨어 추상화 (GPU/NPU)

벤더 종속성 제거 | 통합 제어 | 비용 효율 극대화

올인원 서빙 및 트래픽 제어

원클릭
분산 배포 및 관리

무중단
Failover&Failback

지능형 트래픽 라우팅
및 A/B 테스트

손쉬운 사용성 | 무중단 운영 | 제어 편의성

AURDA

단일 플랫폼

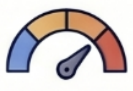
MCP 기반 Agentic ChatOps

AI팀 Namespace에
A100 GPU 2장 할당해줘.

네, AI팀 Namespace에 A100 GPU
2장 할당하겠습니다.
{allocate_gpu 실행}

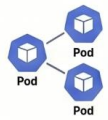
사내 에이전트 연동하여 자연어 요청 | 운영 편의성

End-to-End 통합 모니터링



GPU/NPU
온도

TPS
(초당토큰수)



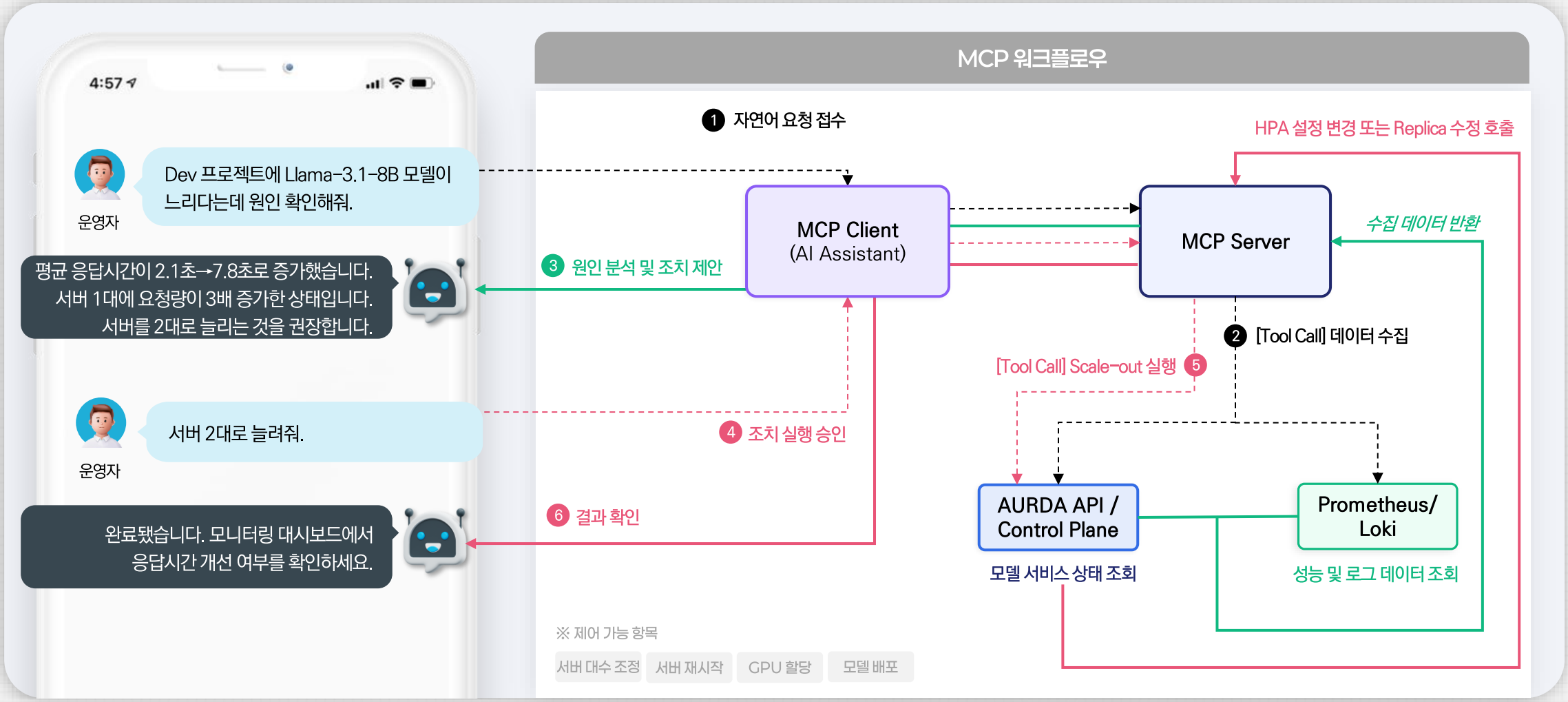
컨테이너/노드
관리

보안/감사
추적

전 구간 단일 대시보드에서 추적 | 신속한 원인 파악

핵심 특장점 : MCP 기반 Agentic 인프라 제어

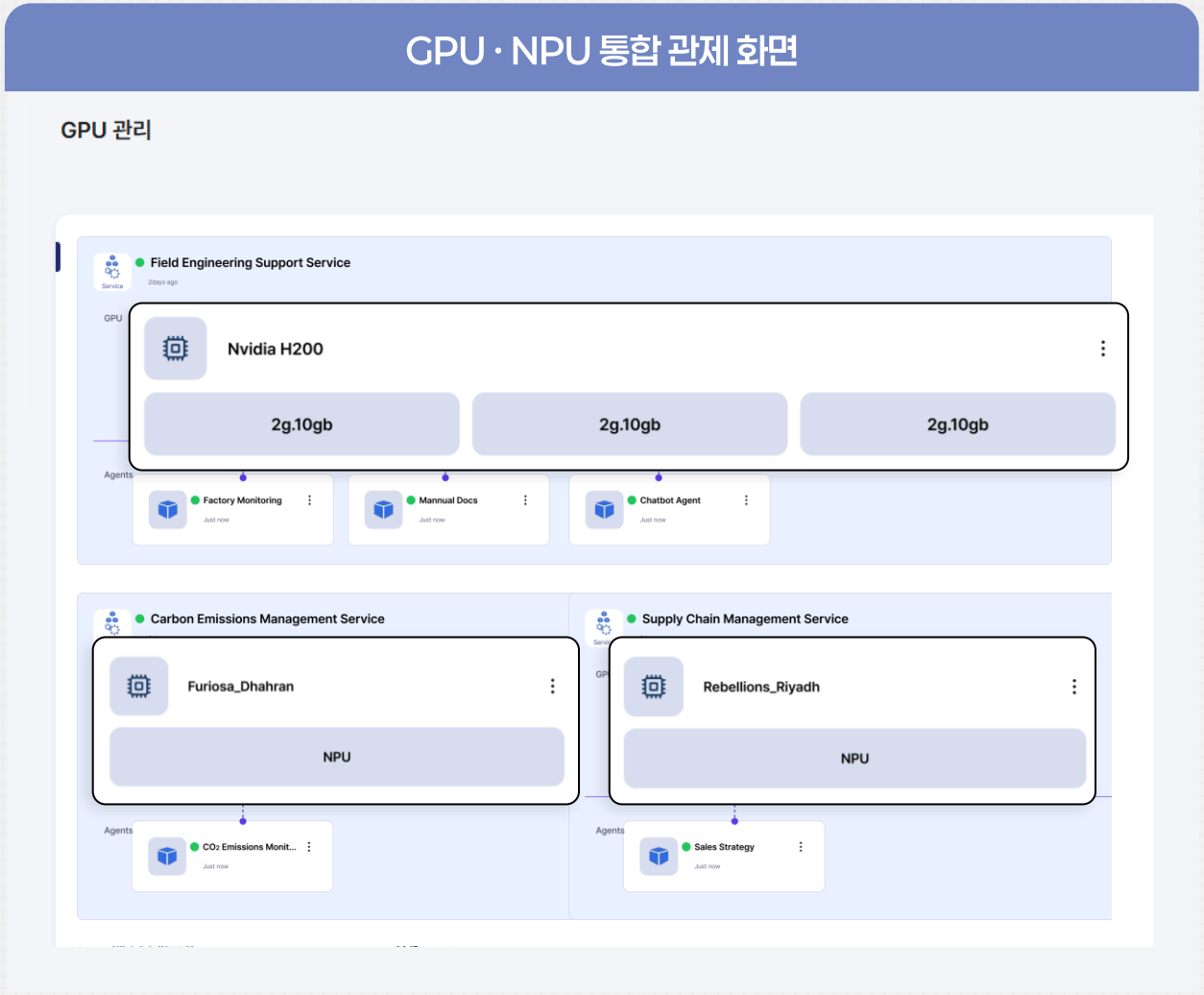
자연어 기반 인프라 제어로 장애 대응 시간을 단축하고, 반복적인 운영 업무를 자동화합니다.



핵심 특장점 : 이기종 HW 통합 관제 (GPU+NPU)

NVIDIA GPU와 국산 NPU를 단일 플랫폼에서 통합 관리하여 벤더 종속 없이 최적의 AI 인프라를 운영합니다.

GPU · NPU 통합 관제 화면



도입 효과

벤더 종속 탈피

특정 클라우드·제조사에 묶이지 않고
GPU·NPU를 자유롭게 혼합 운용

비용 효율 극대화

추론 워크로드에 NPU 활용 시 동등 부하
대비 전력 최대 35~70% 절감

단일 콘솔 통합 제어

이기종 장비가 추가되어도 기존
UI에서 즉시 인식 및 관리

무중단 스케일아웃

GPU·NPU 노드 추가 시 자동 인식 및 편입
서비스 중단 없이 즉시 자원 확장

Key Features : 역할별 운영 흐름

AI 인프라 구축부터 Agent 서비스 운영까지, AURDA 에서 모두 완성됩니다.

인프라 운영자

인프라 및 기반 설정

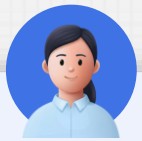
클러스터 설치
- K8s 노드 등록
- 폐쇄망 배포

자원 할당
- GPU/NPU 풀링
- MIG 격리 설정

GitOps 배포
- ArgoCD/Helm
- 형상 관리

보안 설정
- IAM/RBAC
- 네임스페이스 격리

안전하게 격리된 AI 인프라 환경



AI 플랫폼 관리자

AI 런타임 환경 구성

모델 서빙
- vLLM 설정
- 추론 엔진 배포

트레픽 제어
- 라우팅/Failover
- A/B 테스트

추론 최적화
- NPU 활용
- MIG 템플릿 적용

관측 설정
- LLM 트레이싱
- TTFT/TPS 수집

운영 가능한 AI 서비스 런타임

AI 개발자

AI 서비스 개발 및 활용

RAG 파이프라인
- 벡터 DB 연동
- 문서 임베딩 설정

AI Agent 생성
- LLM 모델 연결
- 워크플로우 설계

서비스 배포
- 그룹 단위 권한 설정
- 엔드포인트 발행

가드레일 & 모니터링
- 안전성 정책 설정
- 응답 품질 추적

AI 서비스 완성 및 운영

Key Features : GPU 자원 통합 관리 및 할당

다양한 GPU 자원을 통합 관리하고 워크로드에 맞춰 정밀하게 배치함으로써 자원 낭비를 줄이고 운영 효율을 극대화합니다.

MIG 기반 동적 GPU 분할

하나의 GPU를 독립된 인스턴스로 분할, 워크로드별 연산·메모리 물리 격리

GPU Operator 자동 스케줄링

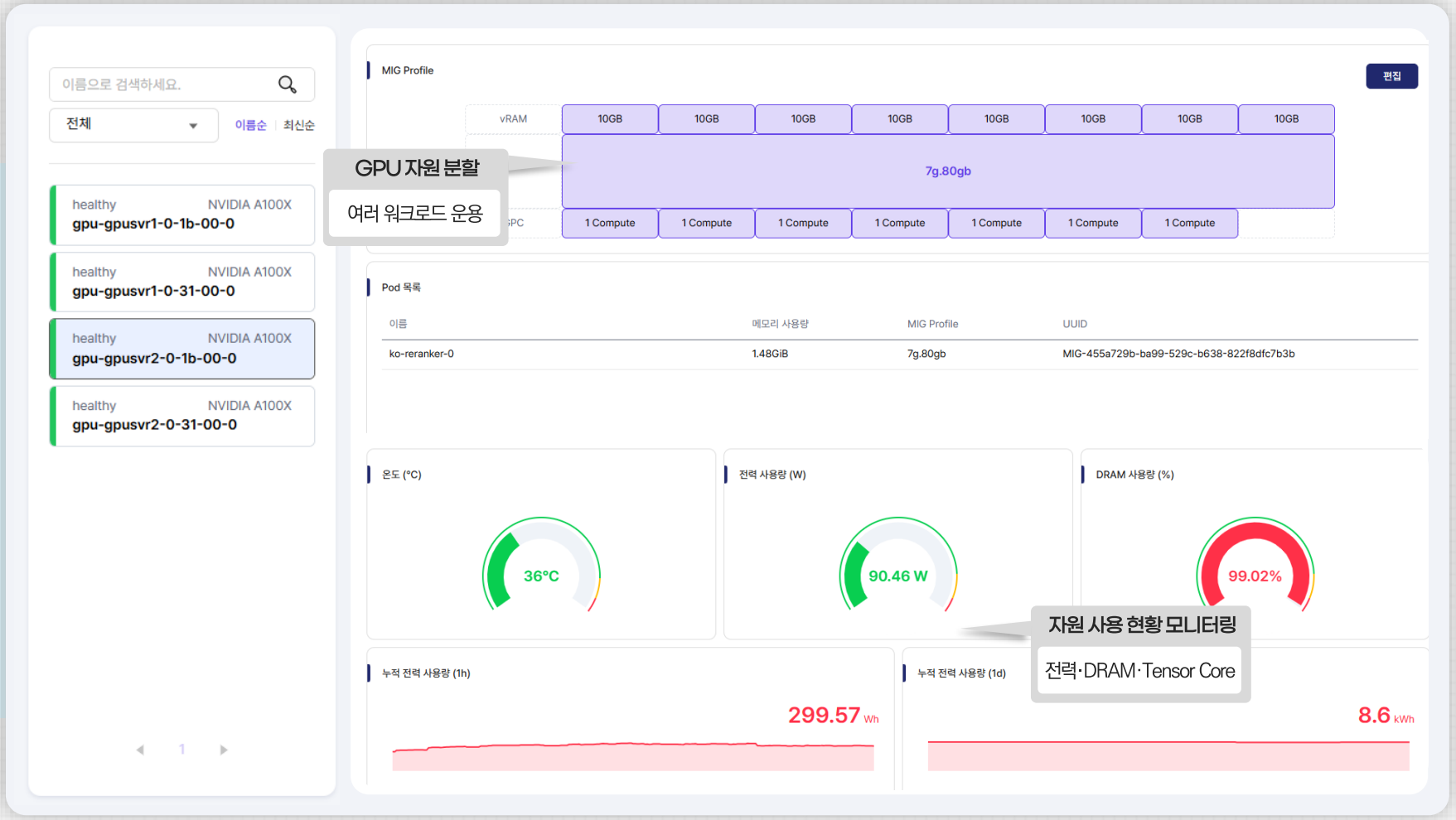
드라이버 설치부터 런타임까지 자동화, 요청 사양에 맞는 최적 노드 즉시 배치

네임스페이스 단위 쿼터 관리

부서·프로젝트별 CPU/Memory/GPU 상한선 설정으로 자원 독점 방지

이기종 하드웨어 추상화

NVIDIA GPU와 NPU를 단일 콘솔에서 혼합 운용 및 제어



Key Features : GitOps 배포

Git 저장소 하나로 인프라와 애플리케이션 배포를 코드로 관리하고, 언제든 안정적으로 되돌릴 수 있습니다.

ArgoCD 기반 선언적 형상 관리

Git에 정의된 상태를 클러스터에 자동 반영, 모든 인프라 구성을 코드로 일관되게 유지

Helm Chart 기반 모듈식 배포

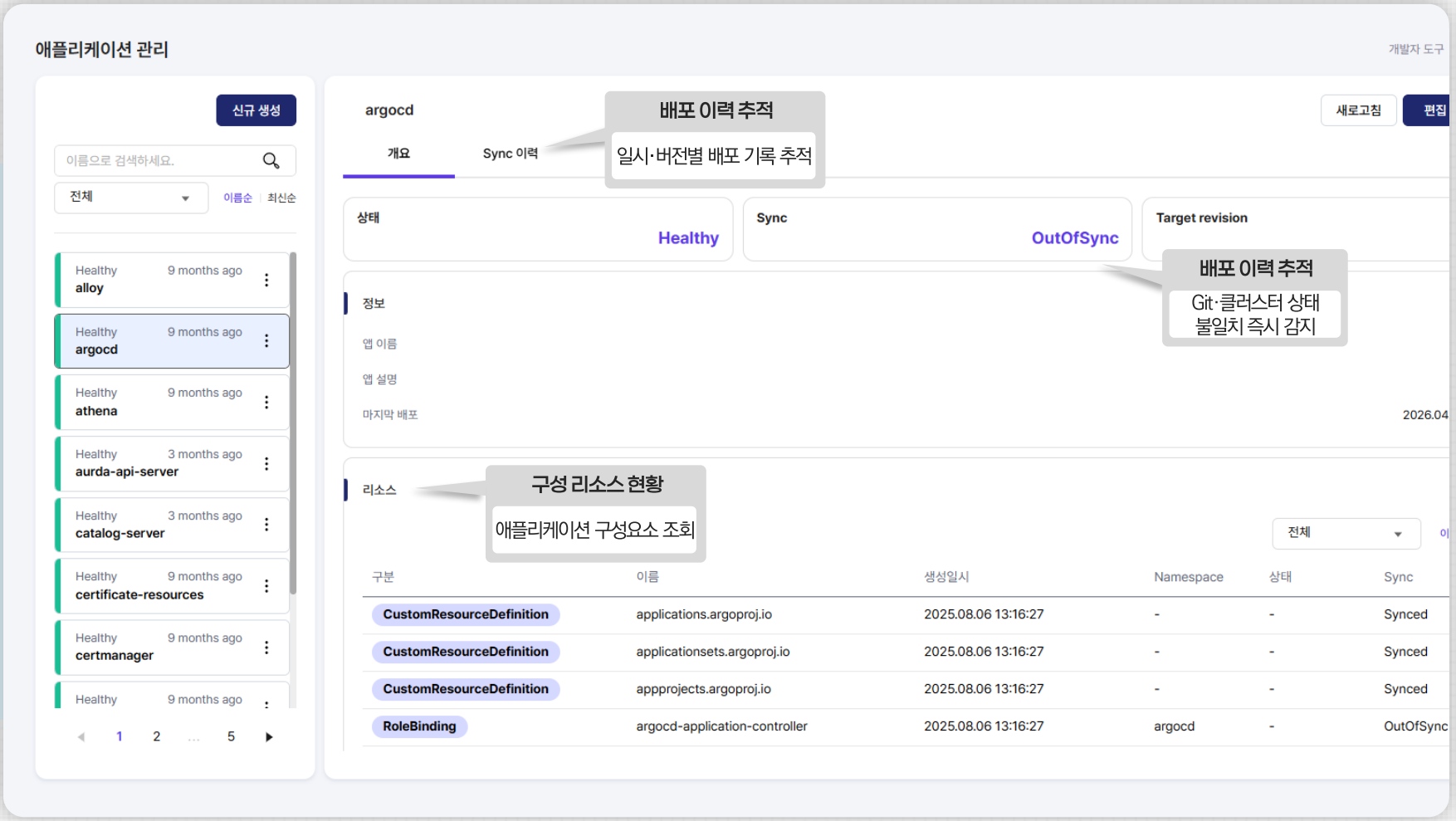
AI 서비스·인프라 구성요소를 패키지로 표준화 하여 환경별 (Dev/Prod) 맞춤 배포 수행

실시간 상태 동기화 및 Drift 감지

Git 정의와 실제 클러스터 상태 차이를 즉시 감지, 자동 복구 또는 운영자 알림으로 형상 강제 유지

배포 이력 관리 및 즉각 롤백

모든 변경 이력을 Git 커밋으로 투명하게 기록, 장애 발생 시 이전 버전으로 즉시 복구



Key Features : 모델 통합 관리

다양한 AI 모델의 정보·실행 환경·GPU 자원·배포 정책을 모델 단위로 체계화하여 운영 복잡도를 줄입니다.

모델 카탈로그 및 버전 관리

상용·오픈소스 모델을 카탈로그화 하여 모델 별 메타데이터·버전·운영 상태를 체계적으로 관리

모델 별 추론 환경 구성

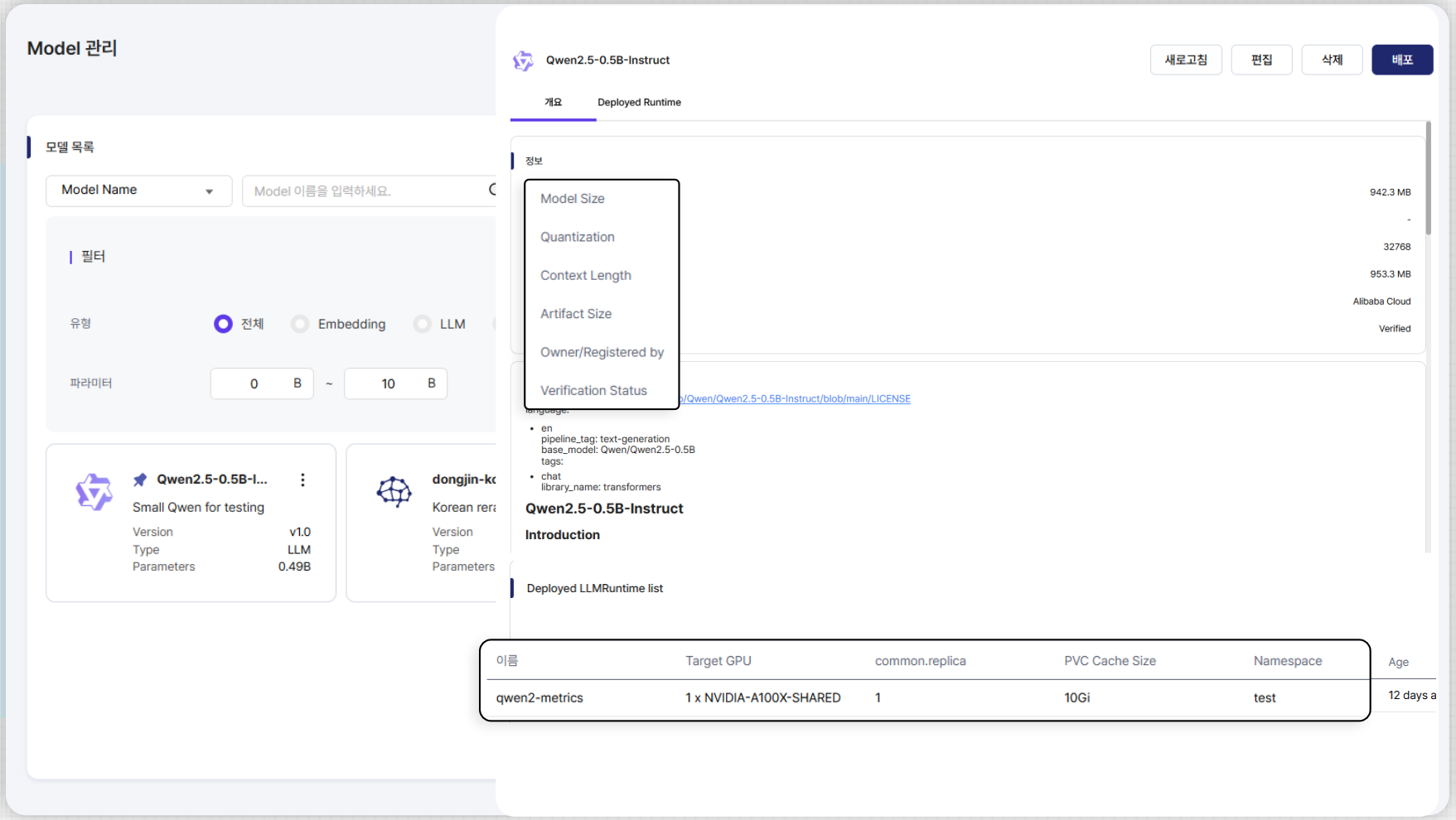
vLLM 등 모델 별 Serving Engine과 추론 설정을 구성해 최적의 실행 환경 제공

모델 저장소 관리

Object Storage 기반으로 대용량 모델 파일을 안전하게 저장하여 외부 연결이 제한된 환경에서도 안정적인 모델 운영 지원

모델 배포 구성 관리

서비스 규모와 처리량에 맞춰 모델의 배포 방식과 복제 구성을 설정하고 일관된 배포 정책으로 관리



Key Features : 고성능 추론 엔진

다양한 LLM을 서비스 환경에 맞게 배포하고, 표준화된 API와 고성능 추론 기능을 통해 안정적인 모델 서비스를 제공합니다.

UI 기반 모델 배포

모델과 실행 환경을 선택해 복잡한 설정 없이 빠르게 모델 서비스를 구성·배포

고성능 추론 처리

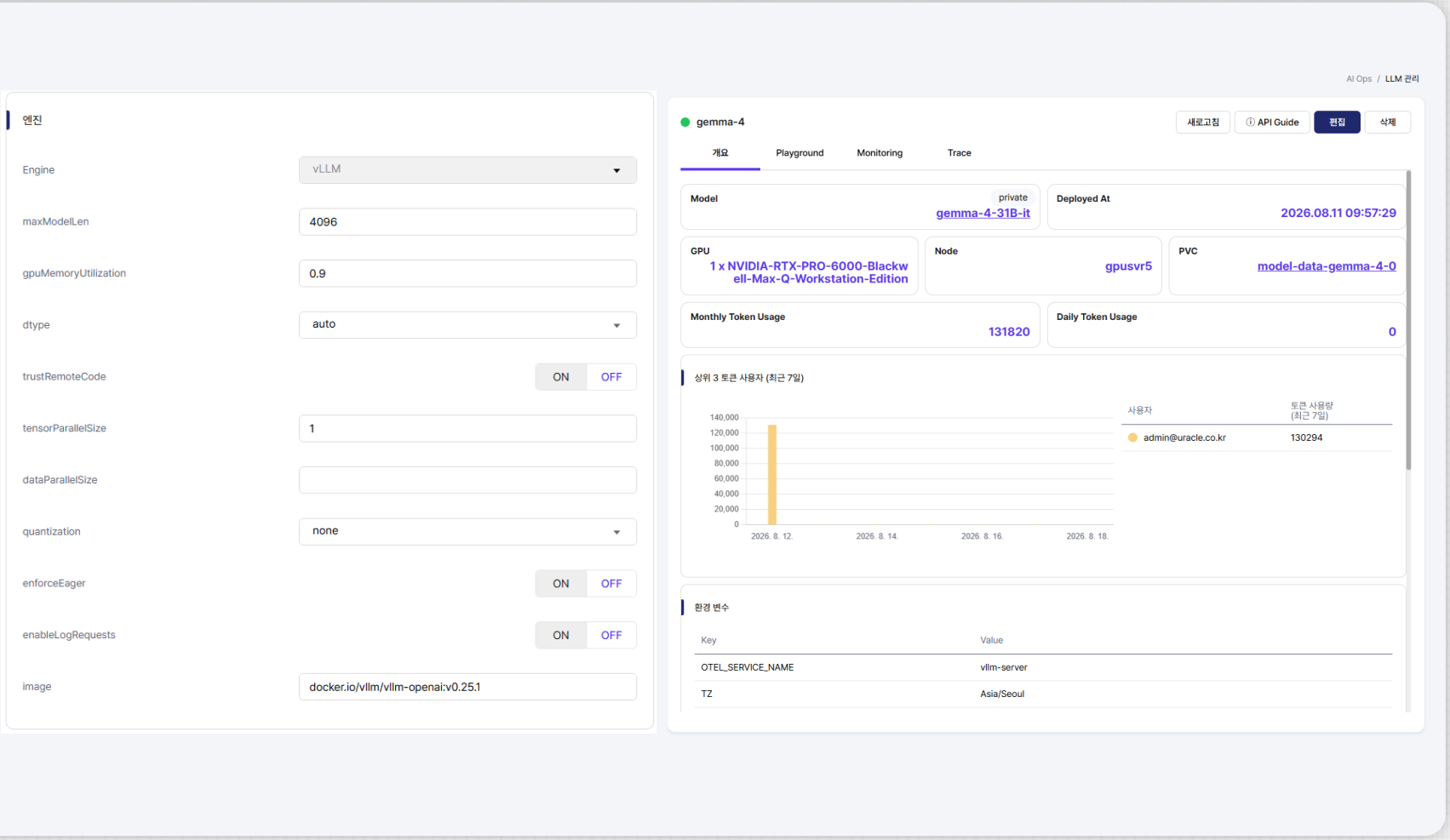
vLLM 기반의 효율적인 요청 처리와 Cache 활용으로 동시 요청 처리량과 추론 성능 향상

모델 별 단일 Endpoint 제공

분산된 모델 인스턴스를 하나의 Endpoint로 통합하여 서비스에서 일관된 방식으로 모델 호출

표준 추론 API 지원

OpenAI Compatible API 기반으로 기존 AI 서비스와 쉽게 연계하고 Streaming 등 다양한 호출 방식 지원



Key Features : 고성능 모델 추론

Multi-GPU·Multi-Node 기반으로 추론 자원을 유연하게 확장하고, 처리 단계별 자원을 최적화하여 대규모 AI 서비스의 안정적인 추론 성능을 확보합니다.

분산 추론 환경 구성

다수의 GPU·Node에 추론 처리를 분산하여 대규모 모델과 높은 요청량에 유연하게 대응

추론 처리 단계 최적화

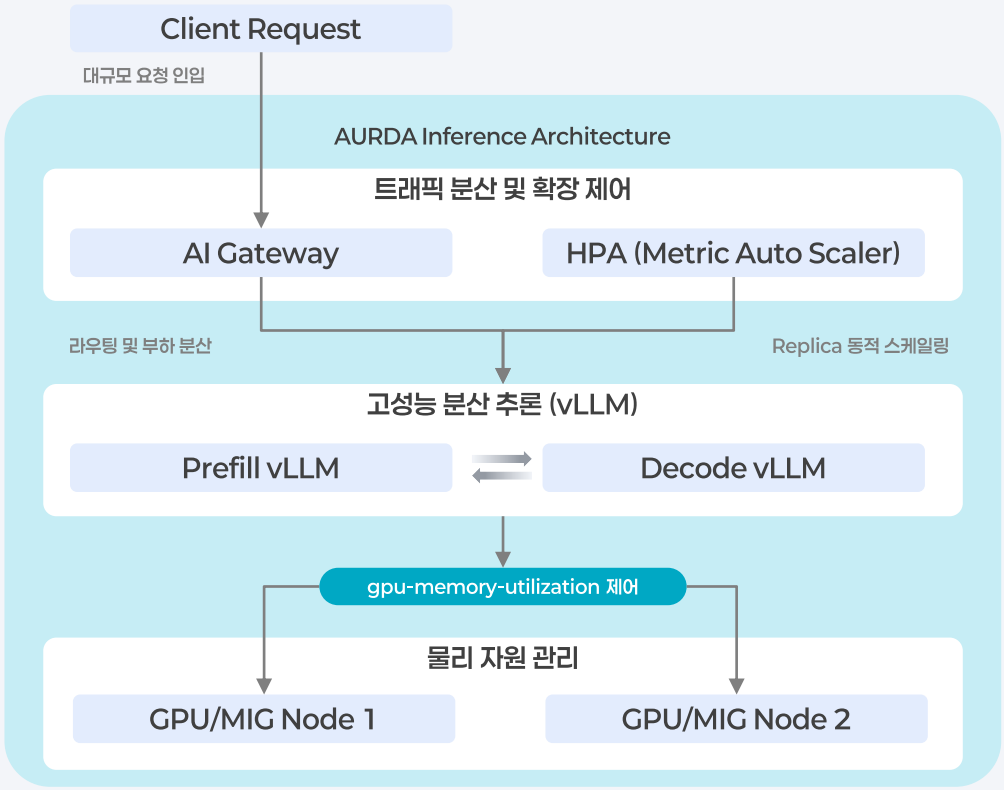
Prefill·Decode 처리 단계를 분리하여 각 단계의 특성에 맞게 GPU 자원을 효율적으로 운영

추론 연산 효율화

KV Cache를 연계·재활용하여 중복 연산을 줄이고 추론 처리 효율과 응답 성능 향상

추론 자원 탄력 확장

처리 단계별 부하에 따라 GPU 자원을 독립적으로 확장하여 변화하는 추론 트래픽에 유연하게 대응



- AI Gateway (인입 제어)
  : 가중치 라우팅 및 Fallback 타겟 자동 재요청 지원.
- Auto-Scaling (탄력적 확장)
  : HPA 메트릭 기반 vLLM 인스턴스 Replica 수 자동 조정.
- vLLM Serving Engine (성능 최적화)
  : TP(Tensor Parallelism)/PP(Pipeline Parallelism) 적용 및 gpu-memory-utilization 기반 운영.
- GPU Operator & MIG (자원 분산)
  : Namespace 단위 GPU 리소스 할당 및 동적 프로파일 관리.

Key Features : 추론 트래픽 제어

추론 요청량과 서비스 상태를 기반으로 트래픽을 제어하고, 장애·과부하 상황에서도 안정적인 모델 서비스를 유지합니다.

트래픽 보호 및 장애 대응

요청량을 정책에 따라 제어하고
Circuit Breaker·Fallback을
통해과부하 및 장애 상황에서도
안정적인 서비스 운영 지원

Load Balancing

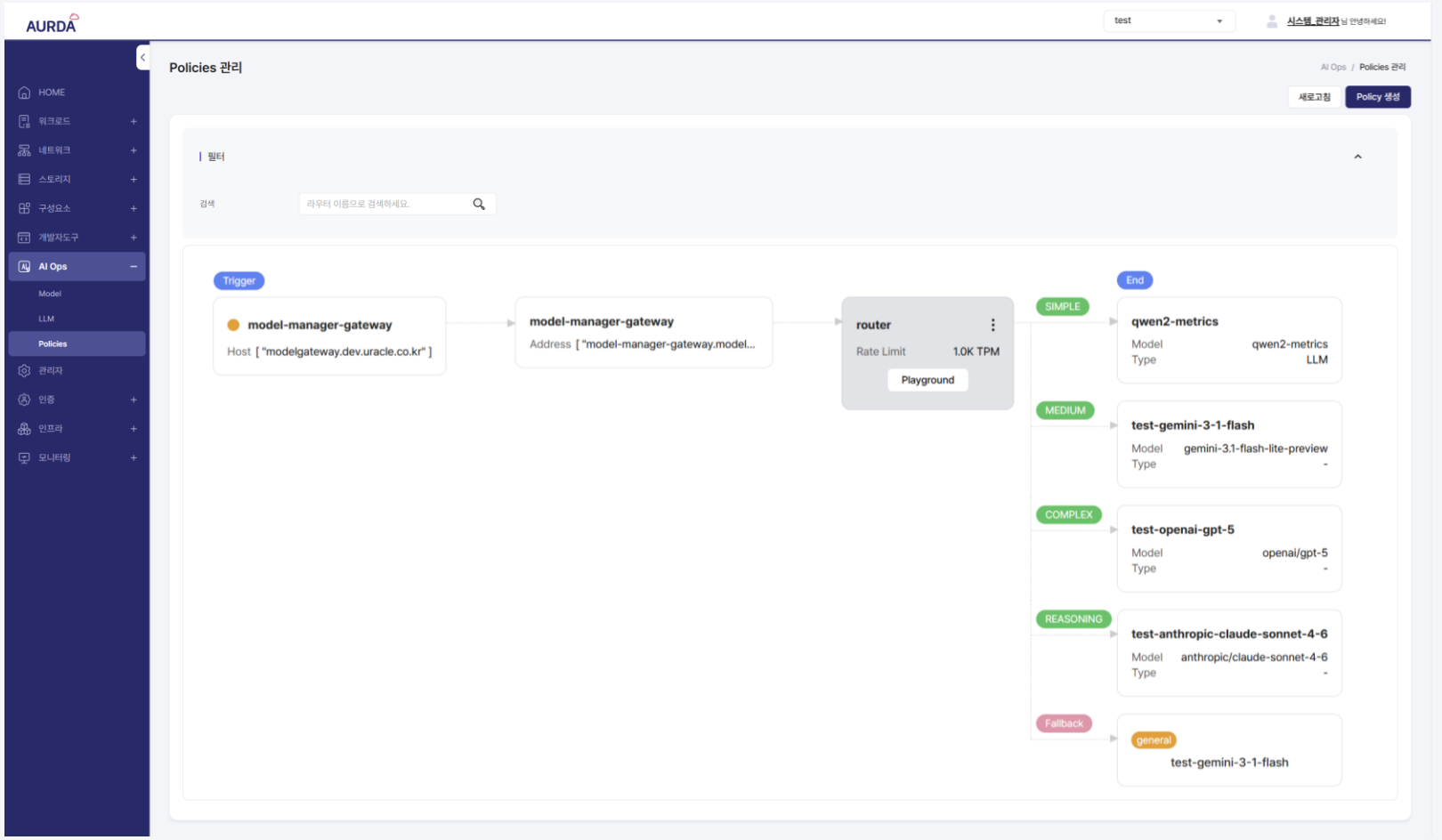
실시간 부하·응답 속도·우선순위
등을 고려하여가용 자원으로
추론 트래픽을 자동 분산

요청량 정책 관리

RPM·TPM·동시 요청 수를
기준으로 서비스별 트래픽 허용
범위와 처리량 제어

우선순위 기반 요청 처리

서비스 중요도와 운영 정책에 따라
요청 처리 우선순위와 분배 기준 설정



※ RPM (Requests Per Minute)
TPM (Tokens Per Minute)

Key Features : 고급 추론 라우팅 전략

Cache·질의 의미·Context 등 요청 특성을 기반으로 최적의 모델과 처리 경로를 선택하여 추론 효율을 높입니다.

Cache 기반 라우팅 최적화

KV·Prefix Cache 활용 상태를 고려해 요청을 분배하여 중복 연산을 줄이고 추론 처리 효율 향상

Semantic 기반 모델 선택 최적화

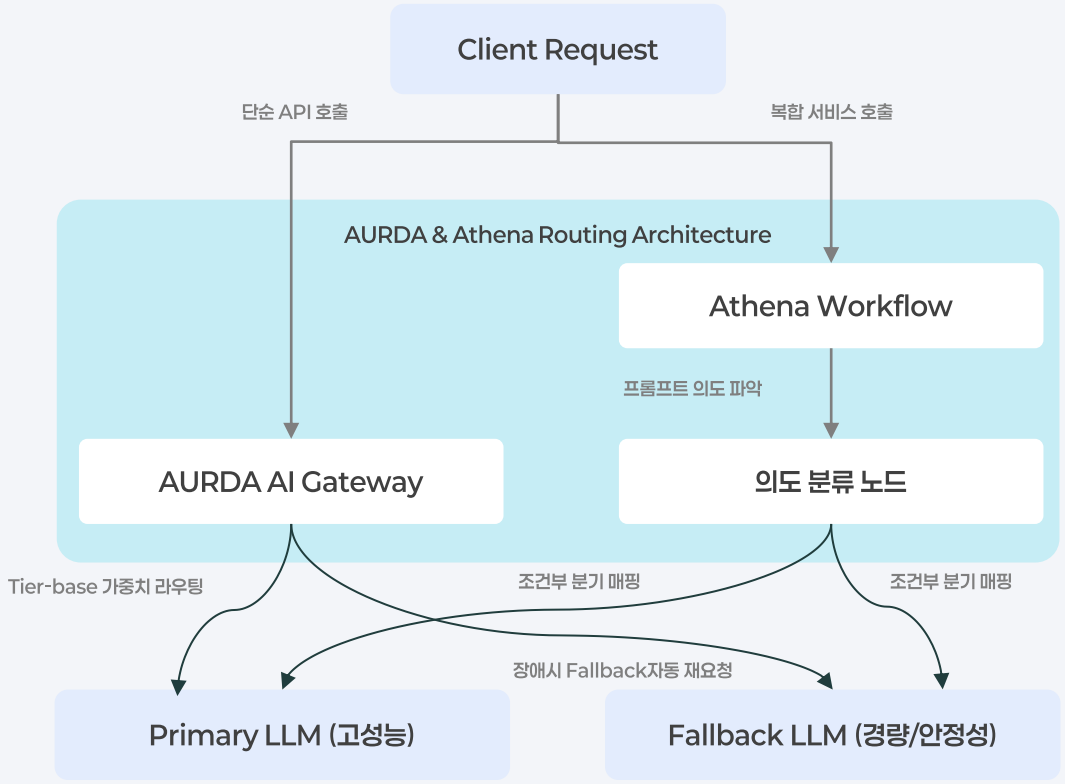
질의의 의미와 업무 영역을 분석하여 요청 특성에 적합한 모델로 자동 연결

Context 기반 요청 분기

Context 크기와 요청 조건에 따라 적합한 모델·Endpoint로 추론 요청을 자동 분기

서비스별 라우팅 정책 적용

서비스·요청별 사전 정의된 조건에 따라 전용 모델·Endpoint로 트래픽을 자동 연결



- AI Gateway 수준의 안정적 트래픽 제어
  : LiteLLM 기반 OpenAI 호환 API 엔드포인트를 통한 Namespace 단위 정책 분리 및 접근 통제
- 무중단 Failover 라우팅
  : Tier-based 가중치 분배 및 모델/노드 장애 시 Fallback 타겟 자동 재요청
- 의도 기반 워크플로우 라우팅
  : Langflow 내 의도 분류 및 조건부 분기(Branch) 노드를 활용한 비즈니스 시나리오별 LLM 매핑

Key Features : 추론 운영 가시성

요청부터 추론 엔진·GPU 자원까지 전 과정을 실시간으로 추적하고, 성능·사용량·오류 정보를 기반으로 안정적인 AI 서비스 운영을 지원합니다.

추론 성능 모니터링

RPS·Latency·TTFT·TPOT 등 주요 지표를 통해 모델별 추론 성능과 응답 상태를 실시간 확인

추론 엔진 상태 분석

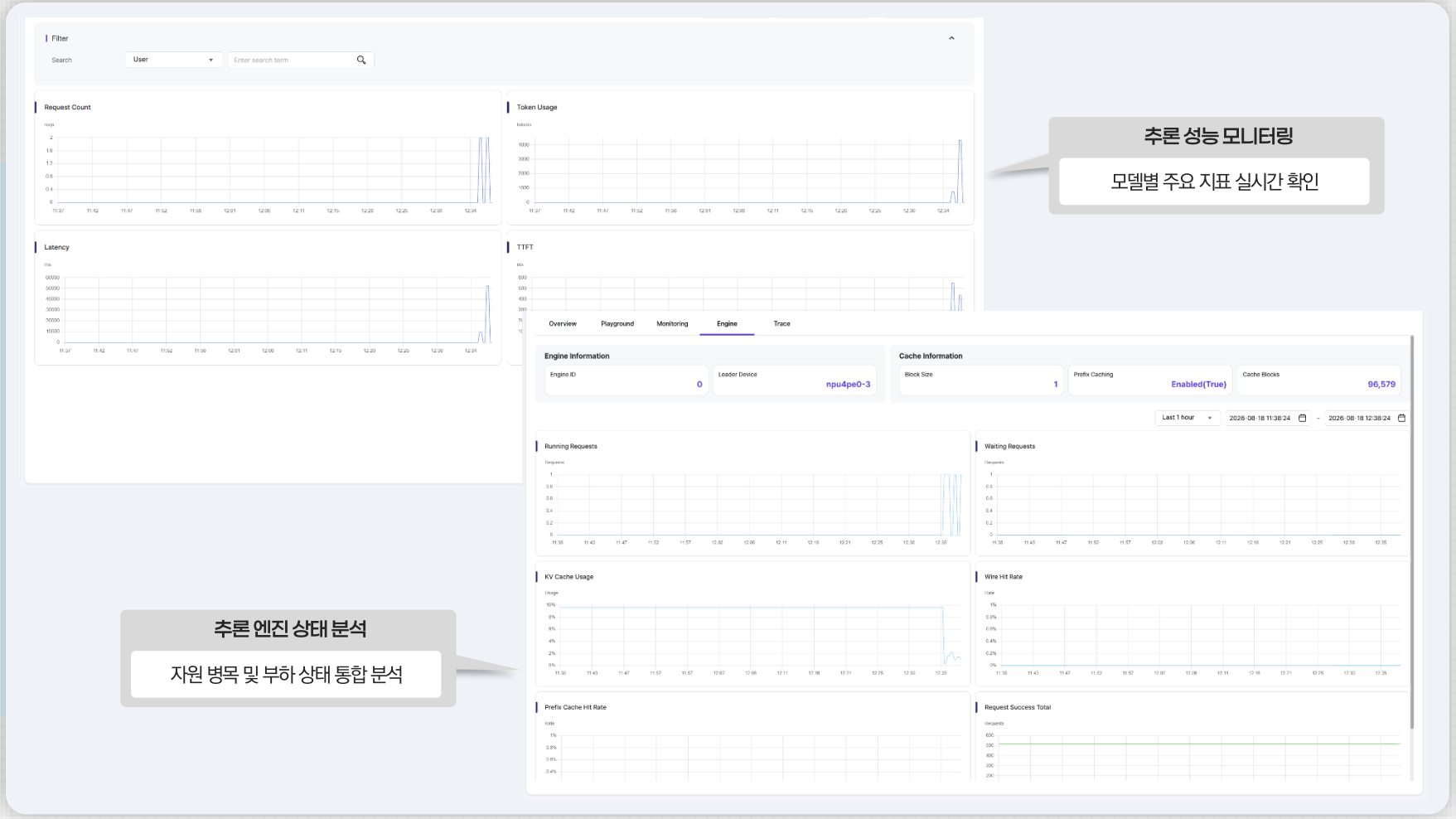
대기·처리 요청과 Queue, Cache·GPU 사용 현황을 기반으로 추론 엔진의 부하와 자원 상태를 통합 분석

End-to-End 요청 추적

Gateway부터 Routing·Inference까지 요청 처리 전 구간을 하나의 Trace로 연결해 지연·오류 발생 구간 확인

운영 이력 및 Audit 관리

사용자·API Key별 호출 이력과 정책·설정 변경 내역을 기록하여 서비스 운영 과정의 추적성과 감사 대응 지원



추론 성능 모니터링

모델별 주요 지표 실시간 확인

추론 엔진 상태 분석

자원 병목 및 부하 상태 통합 분석

Key Features : AI 서비스 모니터링

인프라 자원부터 LLM 추론 전 과정까지 실시간으로 추적하고, 로그와 알림을 통해 서비스 이상 징후에 신속하게 대응합니다.

LLM 추론 전 과정 트레이싱

요청부터 응답까지 단계별 흐름 추적 병목 지점 및 지연 원인 즉시 파악

통합 메트릭 모니터링

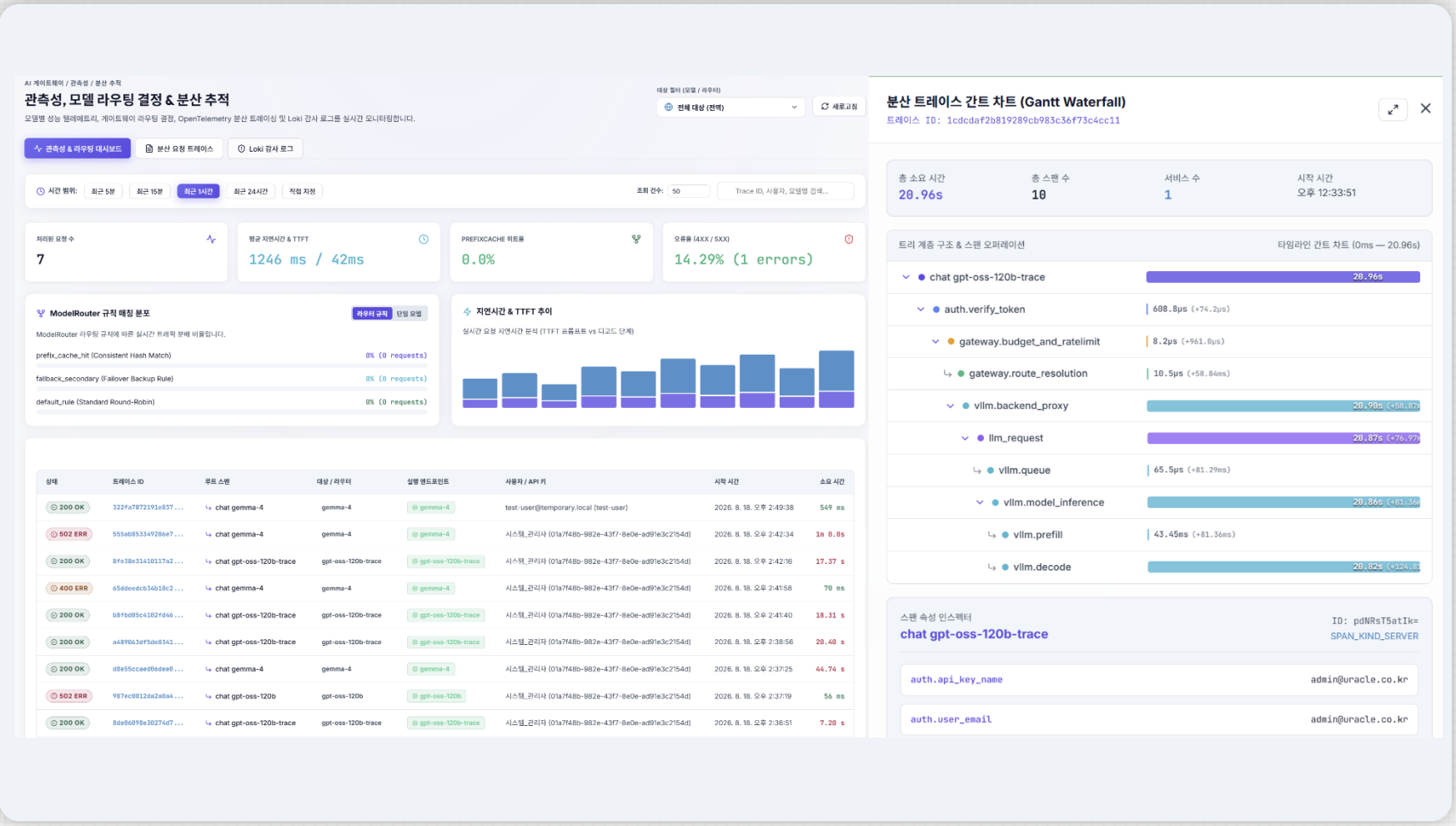
GPU 자원과 LLM 성능 지표를 하나의 대시보드에서 실시간으로 확인하고 서비스 상태를 통합 관리

중앙 집중식 로그 관리

전체 컴포넌트의 로그를 통합 수집·관리하여 장애 발생 시 관련 로그를 빠르게 탐색

임계치 기반 실시간 알림

GPU 상태·오류율 등 주요 지표의 이상 징후를 감지하고 알림 이력을 기반으로 신속한 장애 대응 지원



Key Features : AI 사용량 및 비용 관리

모델·사용자·조직별 AI 사용량과 비용을 통합 분석하고, 예산 정책을 기반으로 효율적인 AI 자원 운영을 지원합니다.

사용자·조직별 사용량 관리

사용자·팀별 모델 호출 및 Token 사용량을 집계하여 조직별 AI 서비스 이용 현황 확인

모델별 사용량 분석

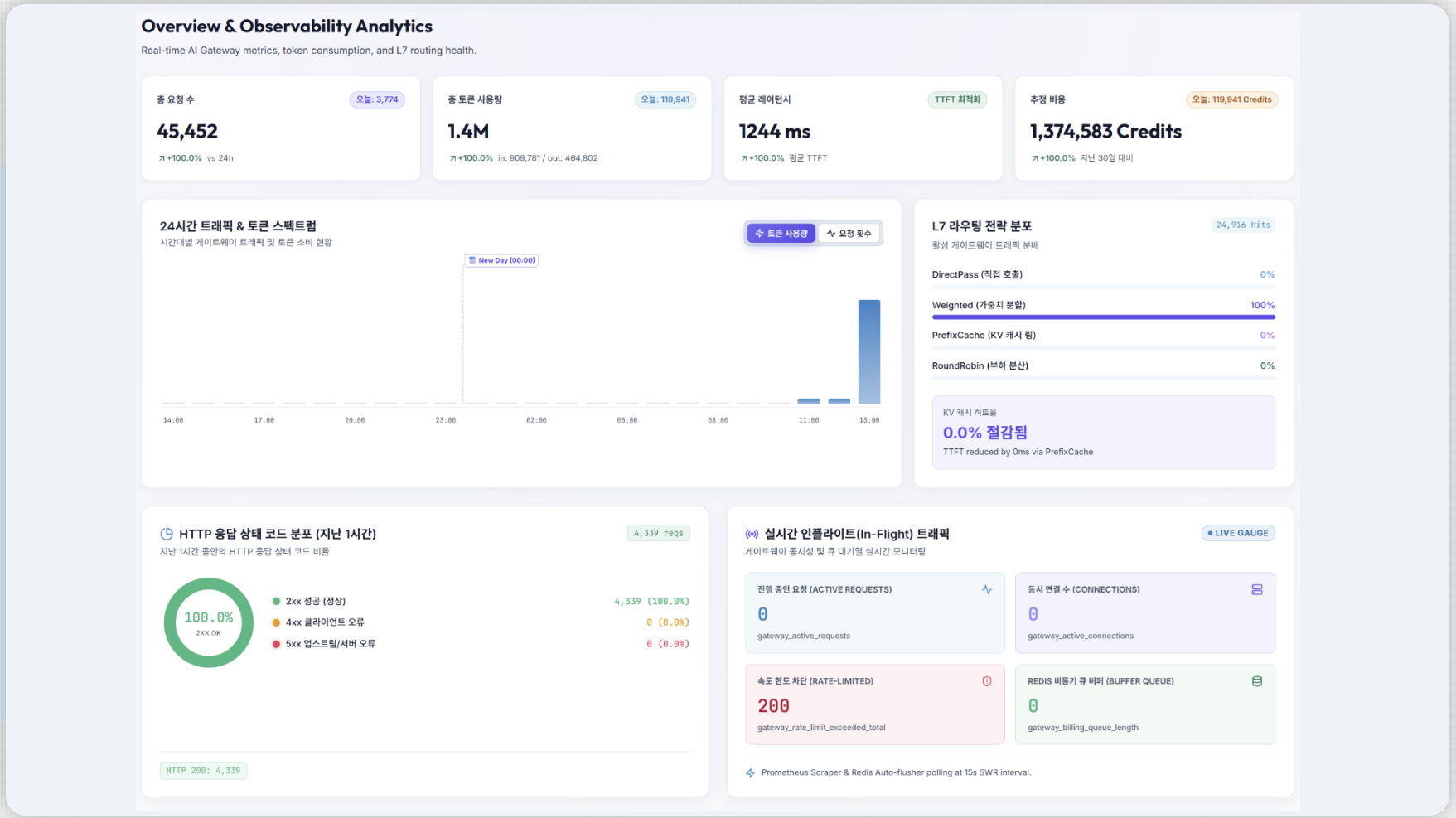
모델별 Request·Input/ Output Token 사용량을 분석해 서비스와 모델 별 활용 현황 비교

AI 비용 추적

Token 사용량을 기준으로 모델 별 비용을 산정하고 서비스·조직별 AI 비용 현황을 통합 관리

예산 및 사용 한도 관리

조직·서비스별 Budget Policy와 사용 현황을 관리해 예산 범위 내에서 AI 자원을 체계적으로 운영



Key Features : 보안 및 접근 제어

사용자·조직·API 단위의 접근 권한과 서비스 정책을 통합 관리하여 AI 모델과 인프라를 안전하게 운영합니다.

역할 기반 접근 제어 (RBAC)

사용자·조직별 역할에 따라 모델과 리소스에 대한 접근 권한을 세분화하여 관리

기업 계정 및 API 인증 연계

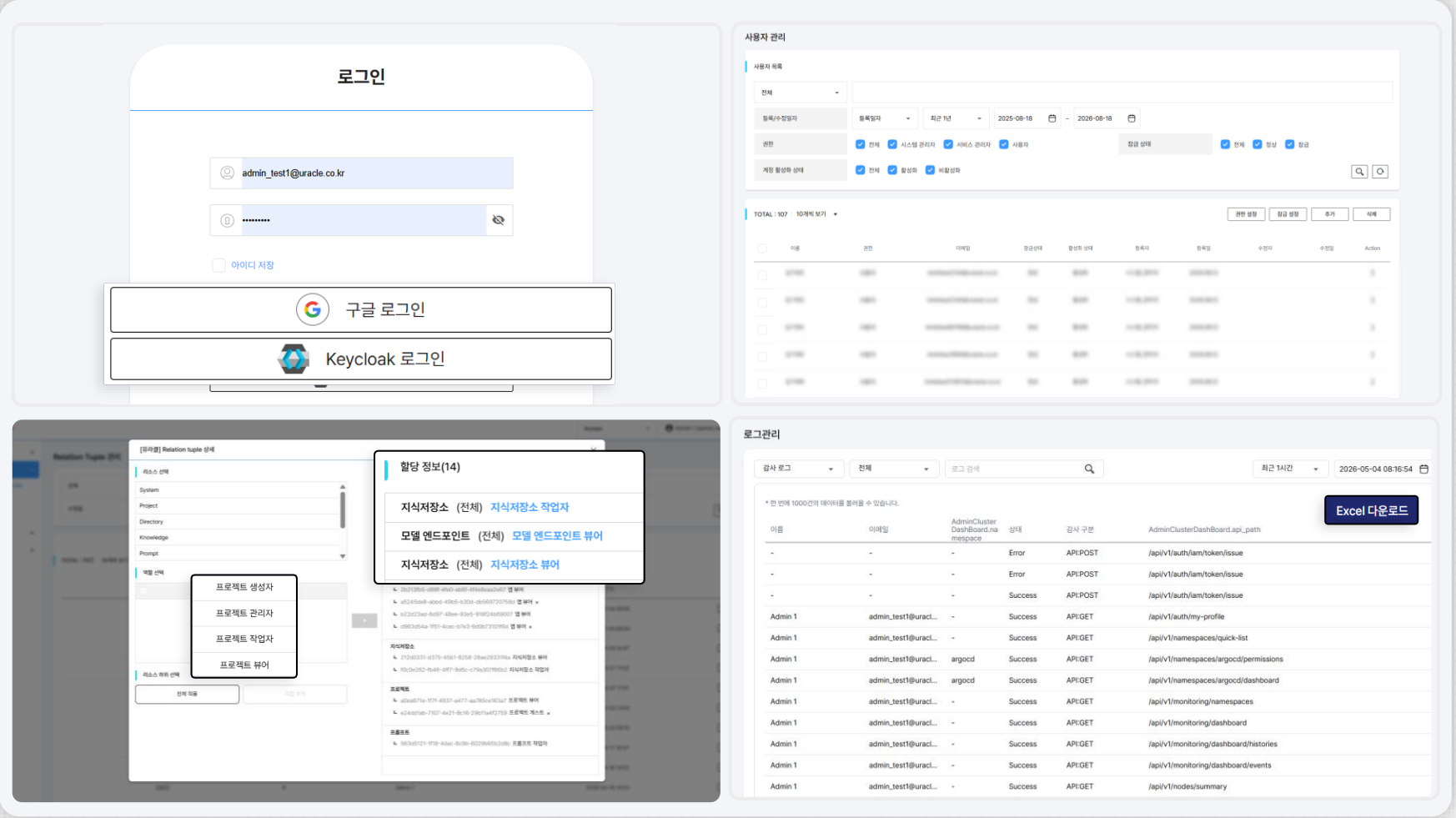
SSO·외부 계정 체계와 API Key 기반 인증을 지원해 기존 기업 인증 환경과 유연하게 연계

모델 접근 및 사용 정책 관리

사용자·팀별 모델 접근 권한과 요청 한도를 설정하여 서비스별 AI 자원 사용을 정책 기반으로 통제

보안 이벤트 및 변경 이력 관리

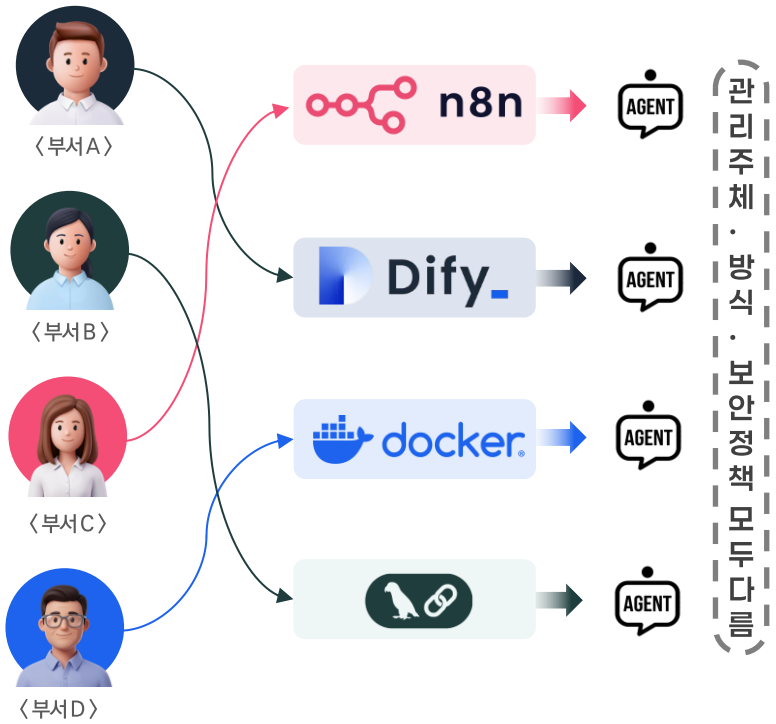
로그인·권한 변경·정책 변경 등 주요 이벤트를 기록하여 운영 이력과 보안 감사 체계 지원



Our Roadmap : Universal AI Runtime

N8N, Docker, 레거시 Agent 등 이기종 AI 서비스를 Athena·AURDA 단일 환경에서 실행·제어합니다.

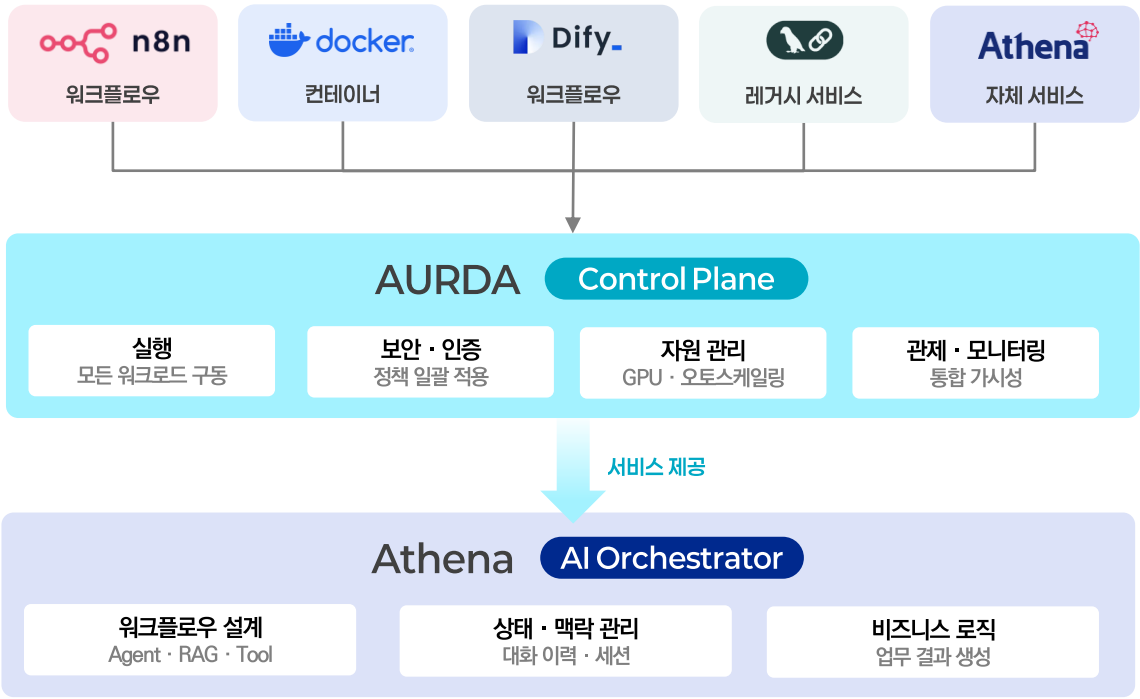
이기종 AI 환경의 사일로화



서비스별 독립 운영으로 보안 정책·자원 관리·모니터링 기준이 상이하여

전사 AI 거버넌스 수립 불가

Control Plane 기반 통합 운영



AURDA Control Plane이 **모든 워크로드의 실행·보안·자원을 단일 기준으로 처리**하고

Athena가 **비즈니스 로직을 오케스트레이션**

Use Cases : 한국평가데이터 - AI 개발플랫폼 구축 (LLMOps + KMS)

표준화된 AI 개발 환경과 사내 지식관리 시스템을 통합 구축하여 전사 AI 서비스 개발·운영 체계를 확립할 예정입니다.

배경

과제별·용역사별로 상이한 아키텍처로 개발된 13개 AI Agent의 유지관리 어려움 증가

고객 주요 과제

- H200 기반 기존 Kubernetes 운영 환경 유지
- AI Agent의 일관된 관리 및 품질 통제
- 표준화된 AI 개발도구와 관리체계 필요
- 사내 지식의 통합·체계적 활용을 위한 KMS 구축

적용 플랫폼: 아테나 / 오르다 / 아테나 코드

구현 서비스

AI 개발플랫폼 (LLMOps)	· 인프라·모델·데이터·Agent를 통합한 표준 AI 개발운영 플랫폼
지식관리 시스템 (KMS)	· 그룹웨어 기안문서·제규정·매뉴얼 등 사내 문서를 자동 수집·벡터화 · 하이브리드 검색 기반 지식 저장소 구축
업무지원 챗봇 3종	· 사내 업무지식 검색 / 고객 상담지원 / 고객 계약현황(DB 연동) / 코드 어시스턴트 (RAG 기반 개발지원)
보안·거버넌스 체계	· RBAC 기반 접근 제어, 개인정보 마스킹, 프롬프트 인젝션 차단, AI 거버넌스 규정 수립 지원

전사 표준 AI 개발플랫폼 기반으로 신규 Agent 개발·배포 체계 일원화

사내 업무 지식 AI 지식화로 직원 문의·지식탐색 시간 단축

코드 어시스턴트 도입으로 개발 생산성 향상 및 보안 취약점 자동 탐지 체계 확보

Use Cases : 코리아크레딧뷰로 – 전사 AI 서비스 플랫폼 구축

약 1TB 규모의 사내 문서를 AI 지식으로 전환하고, OCR 기반 비정형 문서 처리부터 전사 RAG 서비스까지 통합 AI 플랫폼을 구축했습니다.

배경

AX(AI Transformation)를 통한 서비스 경쟁력 강화 및 업무 효율성 향상 필요성 증가

고객 주요 과제

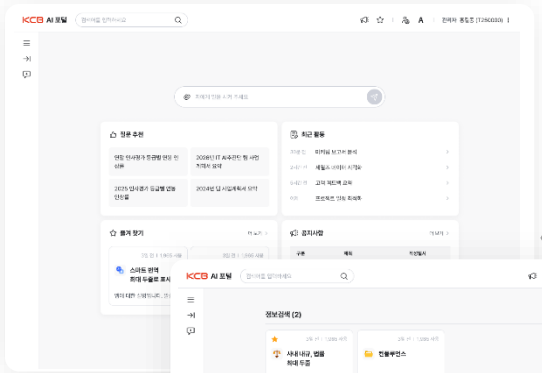
- AI 추진단 4개 운영 중, 전사 직원이 활용할 수 있는 플랫폼 환경 필요
- 약 1TB 규모 사내 문서의 체계적 AI 활용 체계 미흡
- 다양한 비정형 문서 처리를 위한 고정밀 OCR 시스템 필요

적용 플랫폼: 아테나, 오르다, 업스테이지 파서

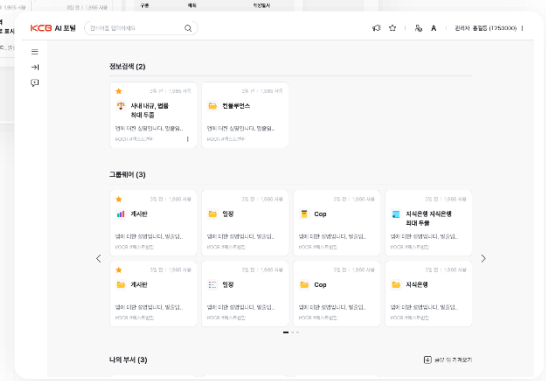
구현 서비스

AI 플랫폼 구축	· 오픈소스 LLM·RAG·데이터 파이프라인 기반 전사 통합 AI 플랫폼 · 망 분리 환경 대응 아키텍처 구성
고정밀 OCR 시스템	· Upstage 상용 OCR 연동 · 스캔본·사진·표·PDF 등 다양한 비정형 문서 인식 및 민감정보 마스킹 처리
지식 검색 · 정보제공 서비스	· 약 1TB 문서 카테고리별 벡터화 · RAG 기반 자연어 질의응답·요약·번역, 근거 문서 함께 제공
관리자 · 운영 환경	· 부서·개인별 AI Agent 생성·공유 · 사용 통계·성능 모니터링 · RBAC 기반 권한 관리 및 DRM 연계

〈사용자 AI포털 메인 UI〉



〈개인 별 템플릿 UI〉



약 1TB 규모 사내 문서 AI 지식화로 전사 검색 정확도 95% 이상 확보

Upstage OCR 연동으로 비정형 문서 인식·처리 자동화 실현

전사 표준 AI 플랫폼 기반 부서별 AI 서비스 자율 확장 환경 구축

Use Cases : 현대건설 - AI 서비스 내재화 플랫폼 구축

RAG 기반 사내 지식 활용부터 개인화 AI 비서까지, 전사 표준 AI 서비스 내재화 플랫폼을 구축 완료했습니다.

배경

AI 활용 업무 개선 트렌드 가속화에 따른
전사 AI 활용 환경 제공 필요성 증가

고객 주요 과제

- 사내 지식의 체계적 AI 활용 체계 부재
- 보안 환경 내 최신 AI 기술 도입 요구
- 현업 니즈 맞춤형 서비스의 지속 발굴

적용 플랫폼 : 아테나, 오르다

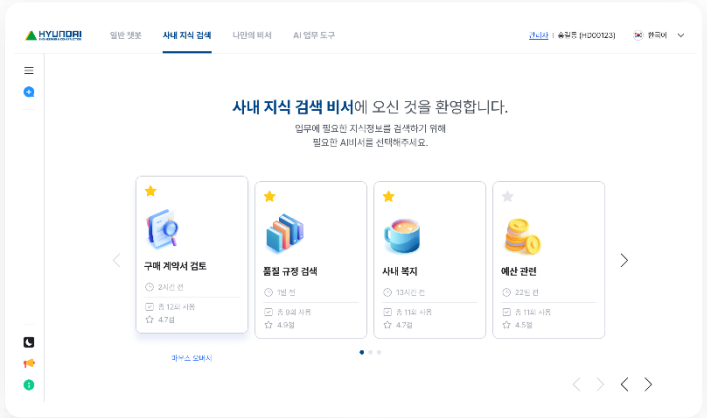
구현 서비스

일반 챗봇 · 사내 지식 검색
- 범용 업무 챗봇 및 구매계약서 검토 /품질규정 검색 등 특화 지식 기반 서비스

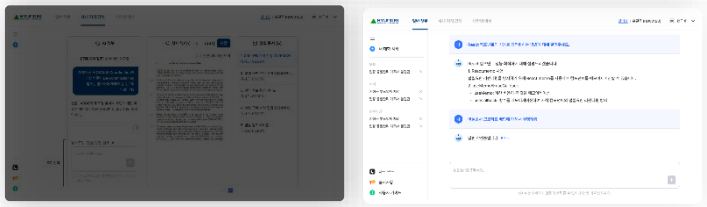
나만의 비서
- 사용자별 개인화 AI 비서 생성
- LLM 모델 선택 및 지식 문서 직접 등록

AI 업무 도구
- 문서번역(DeepL), 회의록 자동 생성
- 문서작성 도구 등 업무 자동화 도구 제공

관리자 시스템
- 사용자·부서별 이용통계
- 금칙어 설정 등 응답 품질 및 운영 관리



〈사내 지식 검색 UI〉



〈일반 챗봇 UI〉

- 임직원 업무 생산성 향상 및 사내 지식 활용 가치 극대화
- SSO·OTP 기반 안전한 AI 서비스 제공
- PC·모바일 통합 서비스로 업무 환경 전반 AI 내재화

Use Cases : 인텔리안테크 – AIOps 플랫폼 및 고객 지원 챗봇 개발

방대한 제품 데이터를 기반으로 24시간 고객 문의에 대응하는 지능형 챗봇과 전사 AI 서비스 확장을 위한 AIOps 플랫폼을 구축했습니다.

배경

제품 사양·모델 비교·호환성 등 다양한 고객 문의에 대응하는 지능형 챗봇 및 전사 AI 서비스 확장을 위한 전략적 AIOps 플랫폼 필요

고객 주요 과제

- 고객 문의 대응 자동화 및 24시간 지원 체계 부재
- 제품 사양·매뉴얼 등 방대한 내부 데이터의 AI 활용 체계 미흡
- 향후 다양한 AI 서비스 확장을 위한 표준 플랫폼 부재

적용 플랫폼: 아테나, 오르다

구현 서비스

AIOps 플랫폼 구축
- Kubernetes 기반 전사 AI 운영 플랫폼 구축
- 클라우드·온프레미스 모두 대응

고객 지원 챗봇
- 제품 사양·모델 비교·재고·수리 현황 등 내부 데이터 기반 답변 제공
- 24시간 한국어 지원

Agent Builder 및 CI/CD
- 노코드 기반 AI 에이전트 설계·배포 환경 제공
- 모델 변경 시 자동 테스트 및 재배포

AI 거버넌스 및 보안
- 정책 기반 데이터 접근 제어·자원 할당·비용 통제
- 개인정보 마스킹 및 RBAC 기반 권한 관리

〈홈페이지 내 고객 챗봇 UI 예시〉

〈CI/CD UI〉

- 제품 전문 챗봇으로 고객 문의 24시간 자동 대응 및 지원 응대 시간 단축
- 기간계 시스템 연동으로 실시간 정보 기반 능동적 고객 응대 실현
- 전사 AIOps 플랫폼 기반 신규 AI 서비스 신속 확장 체계 확보

uracle

H. uracle.co.kr
E. sales@uracle.co.kr
T. 02-3479-4400

2026